Sélection prédictive d'un modèle génératif par le critère AICp

Vincent Vandewalle

Sélection prédictive d'un modèle génératif par le critère AIC_p

Vincent Vandewalle

Laboratoire Paul Painlevé Lille 1 & INRIA
59650 Villeneuve d'Ascq

Résumé

L'obtention de bonnes performances en analyse discriminante est conditionnée par le choix du modèle. Le critère de choix de modèle le plus utilisé dans ce contexte est la validation croisée. Cependant, ce dernier nécessite un temps de calcul important et est sujet à variations. Dans cet article on présente un critère de choix de modèle alternatif, le critère AIC_p. Ce critère cherche à minimiser la déviance de la log vraisemblance évaluée en les étiquettes conditionnellement aux covariables pour un modèle génératif. Après approximation, on obtient un critère théorique dont la pénalité est difficilement calculable en pratique. On propose alors de remplacer cette pénalité par un majorant facile à calculer. La majoration proposée fait intervenir une quantité qu'on appelle dimension prédictive du modèle génératif et dont on précise le sens. Des expériences sur des données réelles montrent l'intérêt du critère proposé.

Abstract

Obtention of good performances in discriminant analysis depends on the model at hand. Cross-validation is the most popular criterion in this context. However cross-validation is both time consuming and subject to variability. In this article we introduce an alternative model choice criterion, the AIC_p criterion. This criterion tries to minimize the deviance of the log-likelihood evaluated on the labels given the covariates for a generative model. After approximation, we obtain an theoretical criterion of which the penalty is hard to compute in practice. We propose to replace the penalty by an upper-bound which is easy to compute. The proposed upper-bound is a quantity which is called predictive dimension of a generative model and which we precise the meaning. Experiments on real data show the proposed criterion usefullness.

mots-clés : modèles génératifs, estimateur du maximum de vraisemblance, classification supervisée, choix de modèle, validation croisée, AIC.

1 Introduction

Les modèles génératifs sont utiles en analyse discriminante. C'est par exemple le cas des modèles de décomposition parcimonieuse de la matrice de variance pour les modèles gaussiens [1]. L'estimation des paramètres de ces modèles est souvent explicite contrairement aux modèles prédictifs de type régression logistique [2] dont l'estimation requiert l'utilisation d'algorithmes itératifs. Ainsi, à temps de calcul égal, davantage de modèles génératifs sont explorés, ce qui a son importance quand le nombre de modèles à explorer est élevé. De plus, si les données sont issues de la distribution postulée, les modèles génératifs possèdent de meilleures propriétés en terme de variance asymptotique que les modèles prédictifs [3].

Cependant, les modèles génératifs font des hypothèses fortes sur la distribution des données. Mettre plusieurs de ces modèles en compétition pour retenir le meilleur est essentiel à l'obtention de bonnes performances. A cette fin, le critère le plus utilisé est la validation croisée [4]. Toutefois, ce dernier peut nécessiter un temps de calcul important et est sujet à variations. D'autres critères de choix de modèle tels que les critères AIC [5] et BIC [6] sont utilisés ; leur calcul est plus rapide mais ils peuvent produire des résultats médiocres en analyse discriminante. Dans cet article on construit un critère de type AIC, le critère AIC_p, qui cherche à évaluer les performances en prédiction du modèle génératif appris. Ce critère est composé de la vraisemblance calculée en les étiquettes conditionnellement aux covariables, et pénalisée par une quantité nouvelle qui s'interprète comme la dimension prédictive du modèle considéré.

Tout d'abord nous rappelons l'utilisation des modèles génératifs en analyse discriminante. Ensuite, nous introduisons un critère de choix de modèle fondé sur une évaluation prédictive des performances du modèle génératif. Enfin, nous montrons le bon comportement de ce critère sur des jeux de données de l'UCI[1] et de Pattern Recognition[2].

2 Rappels sur les modèles génératifs

On dispose de n données $\{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n)\}$ supposées provenir de n réalisations indépendantes du vecteur aléatoire $(\mathbf{X}_1, Z_1)$, où $\mathbf{X}_1$ est le vecteur des covariables à valeurs dans $\mathcal{X}$, un espace mesurable, et où Z_1 représente l'étiquette à valeurs dans $\mathcal{Z} = \{1, \ldots, G\}$, avec G le nombre de classes. Par la suite on notera $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ l'ensemble des covariables observées et $\mathbf{z} = \{z_1, \ldots, z_n\}$ l'ensemble des étiquettes observées. Les modèles génératifs consistent en une modélisation de la vraie distribution de probabilité p par une loi paramétrée $p(\cdot; \theta)$ avec $\theta \in \Theta$ l'espace des paramètres de dimension finie. La décomposition suivante est faite : $p(\mathbf{x}_1, k; \theta) = \pi_k p(\mathbf{x}_1; \theta_k)$ avec π_k la probabilité que $Z_1 = k$ et $p(\cdot; \theta_1)$ la densité de probabilité de $\mathbf{X}_1 | Z_1 = k$.

[1] http ://archive.ics.uci.edu/ml/

[2] http ://www.stats.ox.ac.uk/pib/PRNN/

Le vecteur des paramètres $\theta = (\pi_1, \ldots, \pi_{G-1}, \theta_1, \ldots, \theta_G)$ est estimé par maximum de vraisemblance à partir de $\{\mathbf{x}, \mathbf{z}\}$, et on note cet estimateur $\hat{\theta}$. La règle de classification pour une nouvelle observation $\mathbf{x}_{n+1}$ est déduite par maximum a posteriori : $\hat{z}_{n+1} = \arg\max_{z \in \mathcal{Z}} p(z|\mathbf{x}_{n+1}; \hat{\theta})$. Pour $\mathcal{X} = \mathbb{R}^d$ et si on suppose que $\mathbf{X}|\mathbf{Z} = z \sim \mathcal{N}(\mu_z, \Sigma)$, on retrouve alors l'analyse discriminante linéaire [7].

3 Évaluation prédictive d'un modèle génératif

3.1 Approximation d'un critère idéal par le critère AIC_p

Le critère AIC est construit en approchant la *déviance moyenne* à partir des données à disposition et en utilisant l'hypothèse du bon modèle. Ici, on s'intéresse à la *déviance conditionnelle moyenne*. Supposons qu'on dispose d'un échantillon test $\{\mathbf{x}', \mathbf{z}'\}$ issu de la même distribution que $\{\mathbf{x}, \mathbf{z}\}$, on souhaite alors sélectionner le modèle minimisant :

$$D = E_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{z}'|\mathbf{x}') - \log p(\mathbf{z}'|\mathbf{x}'; \hat{\theta})], \tag{1}$$

où on prend l'espérance sur les réalisations de $\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{z}'$ avec $\hat{\theta}$ obtenu à partir de $\{\mathbf{x}, \mathbf{z}\}$. En utilisant des approximations de type AIC, et en supposant que la distribution des données est issue de la famille paramétrée par θ ($\exists \theta^* \in \Theta / \ \forall\ (\mathbf{x}_1, z_1) \in \mathcal{X} \times \mathcal{Z}\ p(\mathbf{x}_1, z_1) = p(\mathbf{x}_1, z_1; \theta^*)$), on obtient l'approximation suivante :

$$D = \log p(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}) + tr(I - JJ_c^{-1}) + \mathcal{O}_p(\sqrt{n}), \tag{2}$$

où J_c et J sont les matrices d'information de Fisher jointes et marginales évaluées en θ^*. Une critère théorique de choix de modèle pourrait alors être :

$$\log p(\mathbf{z}|\mathbf{x}; \hat{\theta}) - tr(I - JJ_c^{-1}). \tag{3}$$

Cependant, le calcul de $tr(I - JJ_c^{-1})$ est difficile et potentiellement instable puisqu'il nécessite l'évaluation de J qui résulte d'un mélange. Sous l'hypothèse du bon modèle, on peut montrer la majoration $tr(I - JJ_c^{-1}) \leq \bar{\nu}_m$, où $\bar{\nu}_m$ est une quantité qui ne dépend pas de θ^*, qu'on appelle la dimension prédictive du modèle m (voir Section 3.2).

On remplace alors $tr(I - JJ_c^{-1})$ par sa majoration $\bar{\nu}_m$ pour obtenir le critère suivant :

$$\text{AIC}_p(m) = \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}_m) - \bar{\nu}_m. \tag{4}$$

On s'attend à ce que cette pénalité soit trop grande puisqu'il s'agit d'une majoration brutale de la pénalité idéale. Toutefois, dans le cas univarié hétéroscédastique gaussien, si les classes sont confondues, on peut prouver que cette majoration est atteinte. À ce stade du travail, on conjecture que cette borne supérieure pourrait également être atteinte par bien d'autres modèles génératifs quand les classes sont totalement confondues. On suppose en outre que la discrimination doit être réalisée dans le cas de classes peu séparées. Dans cette situation il est en effet crucial de choisir un modèle et l'utilisation de cette pénalité maximale est donc justifiée.

3.2 Définition de la dimension prédictive

Dans la section précédente on a remplacé la pénalité idéale par sa majoration $\bar{\nu}_m$ qualifiée de dimension prédictive du modèle m dont la définition est la suivante :

Définition 1 *Soit m un modèle génératif ayant pour espace des paramètres Θ_m. Soit r un modèle prédictif identifiable ayant pour espace des paramètres Ω_r et vérifiant :*

$$\{\forall \theta \in \Theta_m, \exists \omega \in \Omega_r / \forall (\mathbf{x}_1, z_1) \in \mathcal{X} \times \mathcal{Z}, p(z_1|\mathbf{x}_1;\theta) = p(z_1|\mathbf{x}_1;\omega), \quad (5)$$

$$et\ \forall \omega \in \Omega_r, \exists \theta \in \Theta_m / \forall (\mathbf{x}_1, z_1) \in \mathcal{X} \times \mathcal{Z}, p(z_1|\mathbf{x}_1;\theta) = p(z_1|\mathbf{x}_1;\omega)\}. \quad (6)$$

Alors la dimension prédictive de m est $\bar{\nu}_m = \dim(\Omega_r)$.

On interprète la dimension prédictive comme le nombre de paramètres algébriquement indépendants quand ceux-ci sont estimés en maximisant $p(\mathbf{z}|\mathbf{x};\theta)$ (les problèmes de taille d'échantillon mis à part) ; c'est-à-dire quand les paramètres du modèle génératif sont estimés d'un point de vue prédictif. Ceci justifie le nom de dimension prédictive. Précisons que la preuve de $tr(I - JJ_c^{-1}) \leq \bar{\nu}_m$, s'appuie sur une reparamétrisation du modèle génératif et sur des manipulations de la trace de matrices définies positives. Dans la section suivante nous détaillons le calcul de $\bar{\nu}_m$ dans le cas gaussien.

3.3 Dimension prédictive dans le cas gaussien

Soit $\mathcal{X} = \mathbb{R}^d$ et supposons une distribution gaussienne ϕ conditionnellement à la classe. Prenons la classe G comme classe de référence. On a :

$$\log \frac{\pi_k \phi(\mathbf{x}_1;\mu_k,\Sigma_k)}{\pi_G \phi(\mathbf{x}_1;\mu_G,\Sigma_G)} = \eta_k + \beta_k' \mathbf{x}_1 + \mathbf{x}_1' \Delta_k \mathbf{x}_1 \ \forall k \in \{1,\ldots,G-1\}, \quad (7)$$

avec $\eta_k \in \mathbb{R}$, $\beta_k \in \mathbb{R}^d$ et Δ_k une matrice symétrique de $\mathbb{R}^{d\times d}$. Ainsi le modèle prédictif r vérifiant l'équation (5) est $p(k|\mathbf{x}_1;\omega) = \frac{e^{g(\mathbf{x}_1;\omega_k)}}{1+\sum_{j=1}^{G-1} e^{g(\mathbf{x}_1;\omega_j)}}\ \forall k \in \{1,\ldots,G-1\}$ où $g(\mathbf{x}_1;\omega_k) = \eta_k + \beta_k'\mathbf{x}_1 + \mathbf{x}_1'\Delta_k\mathbf{x}_1$. Ceci correspond à la régression logistique quadratique. On élague ensuite le modèle r en fonction de la paramétrisation de Σ_k choisie pour qu'il vérifie (6) et qu'il soit identifiable. On montre que (η_k, β_k) est libre dans $\mathbb{R}^{d+1}$ quelque soit la paramétrisation de la matrice de variance Σ_k choisie. Il y a donc au moins $\alpha = (G-1)(d+1)$ paramètres libres.

Pour l'obtention de modèles parcimonieux [1], la matrice de variance Σ_k est décomposée en valeurs singulières sous la forme $\Sigma_k = \lambda_k D_k A_k D_k'$, puis des contraintes d'égalité entre les classes pour λ_k, D_k ou A_k sont imposées, comme le fait que Σ_k soit diagonale ($\lambda_k B_k$) ou sphérique ($\lambda_k I$). Après étude des simplifications en fonction des contraintes imposées à Σ_k on parvient à calculer la dimension prédictive pour 9 modèles parcimonieux ; les résultats sont présentés Table 1.

Modèle m	Dimension générative (ν_m)	Dimension prédictive ($\bar{\nu}_m$)
$\lambda DAD'$	$\alpha + d + d(d+1)/2$	α
λB	$\alpha + d + d$	α
λI	$\alpha + d + 1$	α
$\lambda_k D_k A_k D_k'$	$\alpha + d + Gd(d+1)/2$	$\alpha + (G-1)d(d+1)/2$
$\lambda_k B_k$	$\alpha + d + Gd$	$\alpha + (G-1)d$
$\lambda_k DAD'$	$\alpha + d + d(d+1)/2 - 1 + G$	$\alpha + d(d+1)/2 - 1 + (G-1)$
$\lambda_k B$	$\alpha + d + d - 1 + G$	$\alpha + d - 1 + (G-1)$
$\lambda_k I$	$\alpha + d + G$	$\alpha + G - 1$
$\lambda_k DA_k D'$	$\alpha + d + d(d-1)/2 + Gd$	$\alpha + d(d-1)/2 + (G-1)d$

TAB. 1 – Dimension générative et dimension prédictive pour certains modèles parcimonieux ($\alpha = (G-1)(d+1)$)

La plus grande différence entre $\bar{\nu}_m$ (dimension prédictive) et ν_m (dimension du modèle génératif) est obtenue pour $d >> G$ quand on considère le modèle $\lambda DAD'$. Dans ce cas, ν_m est quadratique en d alors que $\bar{\nu}_m$ est linéaire en d. Ceci explique la robustesse de l'analyse discriminante linéaire puisqu'un grand nombre de paramètres est estimé, mais seulement un petit nombre de combinaisons d'entre-eux prend part à l'estimation de la distribution conditionnelle. Remarquons au passage que les modèles λI, λB et $\lambda DAD'$ ont la même dimension prédictive, tandis que leur dimension générative est différente. La critère AIC_p risque donc de favoriser le modèle $\lambda DAD'$ par rapport aux modèles λB et λI. C'est certainement une conséquence de la majoration brutale de la pénalité idéale.

4 Expérimentations

On compare AIC, BIC, la 3-fold validation croisée (CV3), la 10-fold validation croisée (CV10) et AIC_p sur des benchmarks disponibles sur le site de l'UCI et Pattern Recognition. Quand un échantillon test est fourni, on l'utilise pour évaluer l'erreur produite par le modèle sélectionné. Si ce n'est pas le cas, on génère aléatoirement 100 jeux d'apprentissage/test de tailles n et $ntest$, et on moyenne l'erreur produite. Les modèles $\lambda DAD'$, λB, λI, $\lambda_k D_k A_k D_k'$, $\lambda_k B_k$ et $\lambda_k I$ sont mis en compétition.

On remarque Table 2 le bon comportement du critère AIC_p. Du fait de la variabilité dans l'estimation de l'erreur sur ces données, la validation croisée ne produit pas systématiquement les meilleurs résultats.

5 Conclusion

On a défini le critère de choix de modèle AIC_p en analyse discriminante focalisé sur la distribution des étiquettes conditionnellement aux covariables. Ce critère est construit

	d	G	n	$ntest$	AIC	BIC	AIC_p	CV3	CV10
Breast Cancer	30	2	400	169	**4,31**	**4,31**	4,52	4,73	4,76
Wine	13	3	89	89	4,89	2,99	**2,24**	2,94	2,72
Pima	7	2	200	332	23,49	**20,18**	**20,18**	24,10	**20,18**
Crab	5	4	100	100	6,57	5,60	**5,49**	5,84	5,84
Iris	4	3	75	75	2,81	2,93	**2,74**	3,86	3,76
Parkinson	22	2	146	49	**12,59**	12,79	12,89	13,44	13,16
Synt	2	2	250	1000	**10,20**	10,90	10,80	**10,20**	**10,20**
Transfusion	4	2	374	374	28,93	28,93	27,76	24,35	**24,34**
Ionosphere	32	2	175	176	**14,87**	**14,87**	16,14	16,34	16,34

TAB. 2 – Erreur produite par les différents critères de choix de modèle

à partir d'un critère idéal dont on a remplacé la pénalité par sa borne supérieure, ce qui a fait apparaître la notion de dimension prédictive d'un modèle génératif. On a calculé cette dimension pour certains modèles gaussiens et on a montré le bon comportement du critère AIC_p sur des données réelles. Dans un travail futur on montrera que la majoration $tr(I - JJ_c^{-1})$ par $\bar{\nu}_m$ peut être atteinte par de nombreux modèles, puis on s'intéressera au calcul de la dimension prédictive pour d'autres modèles que les modèles gaussiens.

Bibliographie

[1] Bensmail, H. and Celeux, G. (1996) *Regularized discriminant analysis*, *Journal of the American Statistical Association*, 91, 1743–1748.

[2] Anderson, J. A. (1982) *Logistic discrimination*, *Handbook of Statistics*, Vol. 2, 169–191, P. R. Krishnaiah and L. Kanal (Eds.), Amsterdam : North-Holland.

[3] O'Neill, T. (1980) *The General Distribution of the Error Rate of a Classification Procedure with Application to Logistic Regression Discrimination*, *Journal of the American Statistical Association*, 75, 154–160.

[4] Stone, M. (1974) *Cross-validatory choice and assessment of statistical predictions*, *J. Roy. Stat. Soc.*, 36, 111–147.

[5] Akaike, H. (1974) *Information theory and an extension of the maximum likelihood principle*, *Second International Symposium on Information Theory*, 267–281.

[6] Schwarz, G. (1978) *Estimating the dimension of a model*, *Annals of Statistics*, 6, 461–464.

[7] Fisher, R. A. (1936) *The use of multiple measurements in taxonomic problems*, *Annals of Eugenics*, 7, 179–188.